DOCUMENT RESUME

ED 317 586 TM 014 638

AUTHOR Breland, Hunter M.; Lytle, Eldon G.
TITLE Computer-Assisted Writing Skill Assessment Using WordMAP (TM).
PUB DATE Apr 90
NOTE 19p.; Paper presented at the Annual Meetings of the American Educational Research Association (Boston, MA, April 16-20, 1990) and the National Council on Measurement in Education (Boston, MA, April 17-19, 1990).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *College Freshmen; *Computer Assisted Testing; Educational Assessment; Essay Tests; Higher Education; Holistic Evaluation; Predictive Measurement; *Writing Evaluation; Writing Skills
IDENTIFIERS *Software Evaluation; *WordMAP Computer Program

ABSTRACT

The utility of computer analysis in the assessment of written products was studied using the WordMAP software package. Data were collected for 92 college freshmen, using: (1) the Test of Standard Written English (TSWE); (2) the English Composition Test of the College Board; (3) verbal and mathematical Scholastic Aptitude Tests; (4) two narrative essays; (5) two expository essays; and (6) two persuasive essays. The variables analyzed by WordMAP were used to predict the score on a single essay and a combined score for the other five essays that three human readers would give. In either situation, the computer could predict the reader's score reasonably well. It is not likely that many institutions will choose to assess writing without using human readers, but the fact that assessment of writing skills can be enhanced through software analysis may make it possible to reduce the amount of labor required, perhaps by using only one reader instead of the two or three usually required. Computer analysis also makes possible a level of feedback to students and teachers that is not possible using human readers alone. Five tables contain data from the study. (SLD)

Computer-Assisted Writing Skill Assessment

Using WordMAP (TM)

Hunter M. Breland
Educational Testing Service

and

Eldon G. Lytle
Linguistic Technologies, Inc.

Paper prepared for presentation at the annual meetings of the American Educational Research Association and the National Council on Measurement in Education, Boston, April 1990

Computer-Assisted Writing Skill Assessment Using WordMAP (TM)

Hunter M. Breland
Educational Testing Service

and

Eldon G. Lytle
Linguistic Technologies, Inc.

Significant progress has been made in recent years in the automated analysis of written products. Several software programs are commercially available and others are in various stages of development. Well-known are software packages like Writer's Workbench (Frase, 1983; Kiefer & Smith, 1983), Grammatik III (Thiesmeyer, 1984; Wampler, Williams, & Walker, 1988), HOMER (Cohen & Lanham, 1984), WANDAH (Von Blum & Cohen, 1984), HBJ Writer, and Rightwriter, which detect a number of features of style and usage--but which have serious limitations (Bowyer, 1989; Gralla, 1988; Hazen, 1986). These simple programs operate primarily by counting and string matching, and they can often be marketed in the form of one to several low-density diskettes. Other programs under development use pattern matching and/or parsing and are consequently much more complex. These systems include IBM's EPISTLE (Heidorn, Jensen, Miller, Byrd, & Chodorow, 1982), a pattern matching program developed at the University of Pittsburgh (Hull, Ball, Fox, Levin, & McCutchen, 1985), and WordMAP (Lytle & Mathews, undated). Of these more complex systems, WordMAP (TM) has the advantage that it has been used extensively in secondary schools, community colleges, universities, and even graduate business schools.

The assessment of writing skill is labor-intensive, especially if any attempt is made to provide examinees with any feedback other than a single

score. Moreover, it is often unreliable because of reader disagreements and the use of one-item assessments consisting of a single essay on a single topic. Computer analysis of written products allows for greater detail in the feedback that can be provided to examinees and can relieve much of the burden on human readers by making detailed judgments unnecessary. Computer analysis can also add to the validity of assessments made by multiple-choice tests and essay tests judged by multiple readers. Finally, computer analysis of writing is not limited to evaluation of one or two samples of writing; as many samples as are available, even lengthy ones, can be analyzed.

To examine the utility of computer analysis in the assessment of written products, we have made use of an extensive data base of writing skill assessments and other assessments collected over a number of years. We have augmented that data based with computer analyses of the same essays that were originally scored by human readers as part of a research study. With such a large array of variables, a number of important questions can be asked: Can computers score essays as well as human readers? Can computers in conjunction with multiple-choice scores of English skills replace readers? If human readers are essential, how many independent readings are needed when computer analyses and multiple-choice test scores are available? To what degree can writing ability be predicted from a single sample of writing?

Data Source

Data used for the project were originally collected and analyzed by Breland et al. (1987). These data consist of College Board scores (Scholastic Aptitude Test, English Composition Test, Test of Standard Written English) and special essays collected for the study. The essays were written

as a part of freshman English composition courses in six different institutions in six different states. Although a total of 270 students completed all assignments (two narrative essays, two expository essays, and two persuasive essays completed at home), a subsample of 92 students was used for the computer analyses. The following variables were available for 92 cases:

The Test of Standard Written English (TSWE). A 30-minute multiple-choice test administered at the time students are applying for college admission.

The English Composition Test (ECT). A 60-minute achievement test in English usually required by the most selective institutions and somewhat more difficult than the TSWE. For some administrations, the ECT includes a 20-minute essay test and, for those administrations, the multiple-choice portion of the test is 40 minutes. Only the multiple-choice portion of the test was used for the present analyses.

The Scholastic Aptitude Test, Verbal Part (SAT-V). A 60-minute multiple-choice test of verbal aptitude.

The Scholastic Aptitude Test, Mathematical Part (SAT-M). A 60-minute multiple-choice test of mathematical aptitude.

Essay #1 Score. The sum of three holistic scores for a 45-minute expository essay, range 3 to 18.

Essay #2 Score. The sum of three holistic scores for a 45-minute expository essay on a second topic, range 3 to 18.

Essay #3 Score. The sum of three holistic scores for a 45-minute narrative essay, range 3 to 18.

Essay #4 Score. The sum of three holistic scores for a 45-minute

narrative essay on a second topic, range 3 to 18.

Essay #5 Score. The sum of three holistic scores for a persuasive essay written first as a draft in class, discussed in a second class period, and rewritten as a take-home assignment. Range 3 to 18.

Essay #6 Score. The sum of three holistic scores for a second persuasive essay topic written in the same way as Essay 5, range 3 to 18.

For Essay #1, the following variables were also available:

Error Rate. A manual count of errors essay conducted independently by two different readers with the two reader counts summed and then divided by the total number of words written.

Word Count. A computer count of the number of words written.

Paragraph Count. A computer count of the number of paragraphs written.

Passive Verb Flags. A computer count of the number of passive verb flags.

To Be Verb Flags. A computer count of the number of to be verb flags.

Subject/Verb Flags. A computer count of subject-verb disagreement flags.

Fuzzy Word Flags. A computer count of fuzzy (or overused) word usage flags.

Run-on Sentence Flags. A computer count of run-on sentence flags.

Dangler Flags. A computer count of dangling preposition flags.

Spelling Flags. A computer count of spelling flags.

Capitalization Flags. A computer count of capitalization flags.

Punctuation Flags. A computer count of punctuation flags.

Flags Score. A composite score based on all flags.

Marks Score. A composite score based on the number of punctuation mark types used.

WordMAP Composite. A complex composite of all WordMAP variables.

Grammar Flags. A computer count of all grammatical flags

Usage Flags. A computer count of all usage flags.

Style Flags. A computer count of all style flags.

Predicting Holistic Scoring for a Single Essay

Table 1 shows correlations between the Essay #1 score and other available variables. The best correlates of this single essay score are the TSWE and Error Rate (both .60), followed by the ECT (.56), SAT-V (.54), Word Count (.50), Marks Score (.48), Flags Score (.47), and the WordMAP Composite Score (.46). Table 1 also shows that all of these variables correlate better with the Essay #1 score than does SAT-M (.36), but it is interesting to observe that even SAT-M is a useful predictor of writing skill. The surprise of Table 1 is that the count of passive verb flags, style flags, and usage flags all correlate positively with the holistic score for this essay. Apparently, these kinds of variables are not considerable important by the readers of these essays.

Table 2 shows a series of multiple regression analyses in which the Essay #1 score is predicted from selected variables. Variable Set 1 included all multiple-choice scores, but only two of these (TSWE and SAT-V) made a significant contribution to the prediction. The shrunken multiple R of .62 is only slightly greater than the zero-order prediction by TSWE of .60.

Variable Set 2 included all computer-generated scores, and ten of these contributed to the shrunken multiple R of .74. The word count, the count of

usage flags, and the count of fuzzy word flags contributed most to the prediction. Again, the positive beta weights for the usage and passive verb flags are interesting. Variable Set 2 shows, as did Page (1968) many years ago, that a computer can predict reasonably well the scores assigned by human readers of essays. It is of interest to observe also that the multiple R of .74 obtained from computer analysis is larger than the zero-order r of .60 obtained from a manual count of errors in the same essays by two different readers.

Variable Set 3 in Table 2 combined multiple-choice tests and computer-generated scores, and seven of these variables contributed to the prediction. The shrunken multiple of .78 is only slightly larger than that of .74 obtained using the computer scores alone.

Predicting Writing Ability More Generally.

We now turn to a somewhat different type of analysis in which we predict, not the score on a single essay, but the combined scores on five different essays excluding the essay analyzed by WordMAP (TM). In other words, we will show that we can predict a student's ability to write more generally using only multiple-choice scores and a computer analysis of a single 45-minute essay. The five-essay criterion is based on a combination of narrative, expository, and persuasive writing written both in class under timed conditions as well as outside of class without timing following class discussion of the essay assignments. In other words, the five-essay criterion is a pretty good measure of each students writing ability. With an alpha reliability of .88, the five-essay criterion correlates well with a number of variables.

Table 3 shows simple correlations between the five-essay criterion and

available variables. The best correlate of the five-essay criterion obtained was for the TSWE (.72), followed by the ECT (.70), Error Rate (.62), SATV (.58), Flags Score (.47), Marks Score (.44), WordMAP Composite (.42), and Word Count (.40). Once again, note that SAT-M is also a good predictor of writing ability (.39), almost as good as some of the other variables. But it is important to emphasize that Error Rate, Flags Score, Marks Score, WordMAP Composite, and Word Count are based on only a single essay which was not one of the five essays included in the five-essay criterion. Again, it is of interest to note the positive correlations generated by style and usage flags. In other words, style (which is concerned with split infinitives, the use of passive and "to be" verbs, the use of first-person references like "I" or "me," and starting sentences with "and" or "but," and the like) appears not to be considered especially important by readers of college freshmen English papers. The same appears to be true of usage (which is concerned with the use of cliche's, vague, weak, or fuzzy words, slang, and colloquialisms, for example).

Table 4 shows that good predictions of writing ability can be made without the use of human readers. The multiple-choice scores of Variable Set 1 yielded a shrunken multiple R of .73, and the computer-generated variables of Variable Set 2 yielded a shrunken multiple of .66. When both multiple-choice scores and computer-generated scores are combined in Variable Set 3, the shrunken multiple increases to .82.

The use of variables like word count and paragraph count may be viewed by some as "systemically invalid" because feedback of this type to the writers of the essays would not necessarily improve their writing ability (Frederiksen & Collins (1989). Others view such variables as "corruptible"

because knowledge of them could result in faking by examinees. But a case can be made for a count of words written on a timed test since it is a good measure of verbal fluency—an ability not often measured by other tests (Sincoff & Sternberg, 1987).

In Table 5 we introduce the human reader as a predictor of writing ability. As we have noted previously, the Essay #1 that we have analyzed by computer was also read and scored holistically by three different readers. The sum of their scores was the dependent variable in Tables 1 and 2. Now we wish to determine how well these reader scores predict the five-essay criterion, which excludes the Essay #1 score. Variable Set 1 in Table 5 shows that the multiple correlation of these three reader scores predicted the five-essay criterion quite well (R = .74), but not nearly as well as the combination of multiple-choice scores and computer analysis (of the same essay) shown in Table 4 (R = .82).

Variable Set 2 in Table 5 adds the Error Rate, and thus two more human readers, to the prediction. The multiple R of .76 shows that even five human readers of a single essay do not do as well at predicting writing ability as did the combination of multiple-choice and computer scores in Table 4.

Variable Set 3 simulates a common type of writing assessment in which two reader scores are combined with one multiple-choice test score. The second and third readers were chosen because their combined performance was better than other reader combinations. The multiple correlation obtained with Variable Set 3 (R = .80) is comparable to that obtained using multiple-choice scores in combination with computer analysis in Table 4 (R = .82). Replacing the TSWE with the ECT in Variable Set 3 makes no significant difference in the analysis.

Variable Set 4 combines the two human readers with the computer-generated scores. The shrunken multiple R of .77 is almost as high as that obtained in the simulated assessment of Variable Set 3, and it avoids the use of multiple-choice tests.

Variable Set 5 in Table 5 uses all available variables to predict writing ability and shows that only a single reading of the essay is necessary when multiple-choice test scores and computer analysis are combined with human readings. Inclusion of the Third Reader in Variable Set 6 did not increase the multiple correlation beyond the .85 value obtainable with only one reading. Note that SAT-M and the paragraph count are supressor variables.

Conclusions

These results are important because they show that assessments of writing skill can be enhanced through the use of text analysis software. Although it is not likely that many institutions will choose to attempt such assessments without human readers, it will clearly be possible to reduce the amount of labor required--perhaps by using only one reading rather than two or three as is at times the custom.

Equally important is that computer analysis of student essays can provide a level of detail in feedback to students, teachers, and others that is not possible using human readers alone. This kind of feedback has important implications for instruction in English composition. Moreover, computer analysis can provide detailed feedback on many written products, even lengthy ones; a teacher of English will normally provide detailed feedback on only a few brief essays.

Finally, the analysis of free-responses in essay form as a means of assessing writing skill would appear to be a promising alternative to multiple-choice tests, which are viewed by some as having negative consequences for instruction--especially in composition instruction.

References

Bowyer, J. W. (1989). A comparative study of three writing analysis programs. Literary and Linguistic Computing, 4 (2), 90-98.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. and Rock, D. (1987). The Assessment of Writing Skill. College Board Monograph No. 11.

Cohen, M. E., & Lanham, R. A. (1984). HOMER: Teaching style with a microcomputer. In W. Wresh (Ed), The computer in composition instruction. Urbana, IL: National Council of Teachers of English.

Frase, L. T. (1983). The UNIX (TM) Writer's Workbench Software Philosophy. The Bell System Technical Journal, 62, 1883-90.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 18 (9).

Gralla, P. (1988). Grammar checkers: Prose and cons. PC/Computing, October, 146-56.

Hazen, M., et al. (1986). Report on Writer's Workbench and Other Writing Tools. ERIC ED 277015. Chapel Hill, NC: University of North Carolina Microcomputing Support Center.

Heidorn, G. E., Jensen, K., Miller, L. A., & Chodorow, M. S. (1982). The EPISTLE text-critiquing system. IBM Systems Journal, 21, 305-326.

Kiefer, K. E., & Smith, C. R. (1983). Textual analysis with computers: Tests of Bell Laboratories' computer software. Research in the Teaching of English, 17, 201-214.

Lytle, E. G., & Mat ews, N. C. (undated). Field test of the WordMAP (TM) Writing Aids System. Panaca, NV: Lincoln County School District.

Page, E. B. (1968). The use of the computer in analyzing student essays. International Review of Education 14, 210-225.

Sincoff, J. B., & Sternberg, R. J. (1987). Two faces of verbal ability. Intelligence, 11, 263-276.

Thiesmeyer, T. (1984). Teaching with the text checkers. In T. E. Martinez (Ed.), Collected essays on the written word and the word processor. Villanova, PA: Villanova University.

Von Blum, R., & Cohen, M. E. (1984). WANDAH: Writing-aid and author's helper. In W. Wresh (Ed.), The computer in composition instruction. Urbana, IL: National Council of Teachers of English.

Wampler, B. E., Williams, M. P., & Walker, J. Grammatik III Users Guide. San Francisco: Reference Software.

Table 1. Correlations Between Predictor Variables and Essay #1 Score (N = 92)

Predictor Variable	Correlation With Holistic Rating*
Multiple-Choice Scores	
TSWE	.60
ECT	.56
SAT-V	.54
SAT-M	.36
Reader Score	
Error Rate**	- .60
Selected WordMAP Variables	
Word Count	.50
Passive Verb Flags	.15
To Be Verb Flags	.00
Subject/Verb Flags	- .26
Fuzzy Word Flags	- .07
Run-on Sentence Flags	- .16
Dangler Flags	- .29
Spelling Flags	- .25
Capitalization Flags	- .04
Punctuation Flags	.00
WordMAP Composite Scores	
WM Composite	.46
Marks Score	.48
Flags Score	.47
Grammar Flags	- .35
Style Flags	.10
Structure Flags	- .35
Usage Flags	.04

*The sum of three ratings made independently by three different readers.

**The sum of error counts made independently by two different readers divided by the number of words written.

Table 2. Multiple Regression Predictions of Essay #1 Score (N = 92)

Dependent Variable	Predictor Variables	P-Value	beta	R*
Essay #1**	Variable Set 1			
	TSWE	.00	.44	.63 (.62)
	SAT-V	.04	.23	
	Variable Set 2			
	Word Count	.01	.29	.77 (.74)
	WM Composite	.02	.23	
	Flags Score	.07	.19	
	Marks Score	.19	.15	
	Structure Flags	.22	-.10	
	Usage Flags	.02	.28	
	Grammar Flags	.10	-.14	
	Passive Verb Flags	.02	.19	
	Dangler Flags	.14	-.12	
	Fuzzy Word Flags	.02	-.26	
	Variable Set 3			
	TSWE	.03	.34	.80 (.78)
	SAT-V	.02	.27	
	ECT	.22	-.19	
	Word Count	.00	.29	
	WM Composite	.11	.12	
	Flags Score	.01	.24	
	Marks Score	.14	.16	

*Figures in parentheses adjusted for the number of predictor variables.
**The sum of three holistic ratings of Essay 1.

Table 3. Correlations Between Predictor Variables and Writing Ability (N = 92)

Predictor Variables	Correlation With Five Essay Score Sum*
Scores on Multiple-Choice Tests	
TSWE	.72
ECT	.70
SAT-V	.58
SAT-M	.39
Reader Scores	
Error Rate, Essay #1**	-.62
Essay #1 Score	.74
Essay #1, 1st Reader Score	.64
Essay #1, 2nd Reader Score	.68
Essay #1, 3rd Reader Score	.61
Selected WordMAP Variables	
Word Count, Essay #1	.40
Paragraphs, "	.03
Passive Verb Flags, Essay #1	.07
To Be Verb Flags, "	- .05
Subject/Verb Flags, "	- .27
Fuzzy Word Flags, "	.04
Run-on Sentence Flags, "	- .11
Dangler Flags, "	- .24
Spelling Flags, "	- .33
Capitalization Flags, "	- .11
Punctuation Flags, "	- .01
WordMAP Composite Scores	
WM Composite, Essay #1	.42
Marks Score, "	.44
Flags Score, "	.47
Grammar Flags, "	- .25
Style Flags, "	.20
Usage Flags, "	.10

*The sum of 15 reader scores on 5 essays excluding Essay #1.
**The sum of error counts for the Essay #1 made by two different readers divided by the number of words written for this essay.

Table 4. Multiple Regression Predictions of Writing Ability Without Human Readers
(N = 92)

Dependent Variable	Predictor Variables	Predictor Significance (P-value)	beta	Multiple R*
Holistic Sum**	Variable Set 1			
	TSWE	.00	.46	.74 (.73)
	ECT	.04	.03	
	Variable Set 2			
	Word Count	.01	.03	.69 (.66)
	Paragraph Count	.01	- .24	
	WM Composite	.10	.18	
	Flags Score	.00	.33	
	Marks Score	.17	.02	
	Usage Flags	.05	.18	
	Dangler Flags	.18	- .12	
	Variable Set 3			
	TSWE	.00	.45	.84 (.82)
	SAT-V	.01	.02	
	SAT-M	.11	- .01	
	Word Count	.01	.02	
	Paragraph Count	.00	- .22	
	Flags Score	.00	.02	
	Marks Score	.05	.02	
	Usage Flags	.01	.16	

*Figures in parentheses adjusted for the number of predictor variables.
**The sum of five essay scores excluding Essay Score 1.

Table 5. Multiple Regression Predictions of Writing Ability With Human Readers (N = 92)

Dependent Variable	Predictor Variables	P-Value	beta	R*
Holistic Sum**	Variable Set 1			
	1st Reader, Essay #1	.00	.29	.75 (.74)
	2nd Reader, "	.00	.40	
	3rd Reader, "	.09	.17	
	Variable Set 2			
	1st Reader, Essay #1	.09	.17	.78 (.76)
	2nd Reader, "	.00	.33	
	3rd Reader, "	.08	.16	
	Error Rate, "	.00	- .27	
	Variable Set 3			
	2nd Reader, Essay #1	.00	.27	.81 (.80)
	3rd Reader, "	.01	.23	
	TSWE	.00	.46	
	Variable Set 4			
	2nd Reader, Essay #1	.00	.41	.79 (.77)
	3rd Reader, "	.18	.13	
	Word Count, "	.14	.02	
	Paragraphs, "	.01	- .20	
	WM Composite, "	.22	.01	
	Marks Score, "	.26	.01	
	Flags Score, "	.02	.02	
	Usage Flags, "	.07	.14	
	Variable Set 5			
	2nd Reader, Essay #1	.00	.27	.87 (.85)
	TSWE	.06	.21	
	SATV	.08	.12	
	SATM	.12	- .01	
	ECT	.19	.17	
	Word Count, Essay #1	.03	.17	
	Paragraphs, "	.00	- .20	
	WM Composite, "	.30	.10	
	Flags Score, "	.12	.11	
	Marks Score, "	.27	.11	
	Usage Flags, "	.01	.19	

*Figures in parentheses adjusted for the number of predictor variables.
**The sum of 15 holistic scores on 5 different essays, excluding Essay #1.